Conference on Applied Statistics in Agriculture

2003 - 15th Annual Conference Proceedings

# A SIMULATION STUDY OF EXPONENTIAL SEMIV ARlO GRAM ESTIMATION

Edward E. Gbur

Bruce A. Craig

Hao Zhang

## Recommended Citation

Gbur, Edward E.; Craig, Bruce A.; and Zhang, Hao (2003). "A SIMULATION STUDY OF EXPONENTIAL SEMIV ARlO GRAM ESTIMATION," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1179

# A SIMULATION STUDY OF EXPONENTIAL SEMIVARIOGRAM ESTIMATION

Edward E. Gbur[1], Bruce A. Craig[2] and Hao Zhang[3]

[1] Agricultural Statistics Lab, University of Arkansas, Fayetteville, AR
[2] Department of Statistics, Purdue University, West Lafayette, IN
[3] Department of Statistics, Washington State University, Pullman, WA

## Abstract

Incorporating the spatial structure of data from agricultural field experiments into inference procedures has become an important topic in recent years. As part of a larger project to determine whether or not reliable predictions and estimates can be obtained for sample sizes often encountered in traditional field experimentation, this paper focuses on the small sample estimation of the parameters of the exponential semivariogram model. Simulation studies were conducted for both expanding and fixed domains. The results indicate large sample to sample variation in sample and fitted semivariograms, neither of which may be "close" to the true model. Distributions of individual parameter estimators are skewed and highly variable. Empirical coverage levels for large sample confidence intervals for the parameters are well below the nominal level and, contrary to what would be expected, decrease as the sample size increases. The results cast doubt on the success of incorporating spatial structure into traditional field data analyses.

Keywords: exponential semivariogram, simulation, small sample estimation, spatial data

## 1. Introduction

The motivation for this research came from our work as part of the North-Central Regional Project NCR-170 "Research Advances in Agricultural Statistics." Our focus in that project was on small sample spatial problems. In particular, we wanted to investigate whether or not reliable predictions and estimates can be obtained for sample/grid sizes often encountered in traditional agricultural field experimentation. We planned and began to carry out a simulation study with different models and sample sizes. However, we soon realized that this work was generating more questions than answers. This paper summarizes some of this work, focusing on the exponential semivariogram model and estimation of the model parameters.

## 2. Motivating Example

Let Z(x,y) be an isotropic, second-order stationary Gaussian random process defined on a region in two dimensions (x,y); i.e., the mean of Z is constant and the covariance between Z at any two points is a function only of the difference between them. Without loss of generality, assume that E(Z) = 0. As a result of the stationarity assumption,

$$\mathrm{var}(Z(s_1) - Z(s_2)) = 2\,\gamma(s_1 - s_2),$$

where $s_i = (x_i,y_i)$, i = 1,2, are any two points in the region. The function $\gamma(\bullet)$ is called the semivariogram.

To illustrate the basic problem, suppose that Z follows an exponential semivariogram model; i.e.,

$$\gamma(h) = c_0 + c_e(1 - \exp(-h/a_e)), \tag{1}$$

with $c_0 = 1$, $c_e = 4$ and $a_e = 3$ and where h is the distance between any two points. Consider a 10 × 10 grid of points (x,y), spaced one unit apart in both directions, located within the region on which Z is defined. Random samples $Z_1, ..., Z_{100}$ were generated on the grid and the sample semivariograms were obtained using the SAS procedures SIM2D and VARIOGRAM. The sample semivariograms for the first six realizations are shown in Figure 1.

Large sample to sample variability is evident in the figure. Moreover, none of the sample semivariograms seem to mirror the underlying model very well (Rep 6 appears the closest). This figure is especially disconcerting because the realizations were not chosen specifically to illustrate the diversity but rather represent the first six realizations generated. Figure 2 plots the first 100 sample semivariograms. This plot reinforces the fact that there is a large amount of sample variability and that many did not appear to be exponential. Several authors (e.g., Diggle et al., 2002; Webster and Oliver, 1992) have commented on the problem of sample semivariogram variability but none that we found convey the impression of its severity as do Figures 1 and 2.

Faced with the above evidence, we turned to a much more fundamental question. Based on the sample semivariogram from a small sample, can we identify the appropriate model which generated the data? As a first step toward answering this new question, if we assume that the form of the semivariogram model is known, how well can we estimate the parameters? Underlying our investigation of this question was the implicit assumption that a reasonably "good" estimate of the semivariogram model is ultimately necessary for "good" prediction and estimation. Webster and Oliver (1992) note that even though predictions obtained by kriging are fairly stable, their prediction variances and, hence, confidence limits, are sensitive to the semivariogram.

## 3. Previous Work

Investigations of the behavior of a statistical procedure in small samples are usually related to its asymptotic properties. In our problem, two possible scenarios of increasing sample size (number of grid points) arise. First, the distance between adjacent points can be held constant and the overall area of the grid increased. Second, the overall area of the grid can be held constant and the distance between adjacent points reduced. We refer to the former situation as the expanding domain case and the latter as the fixed domain or in-fill case.

Zimmerman and Zimmerman (1991) compared five estimators of the exponential semivariogram parameters based on ordinary least squares, weighted least squares, maximum likelihood, REML and generalized MIVQU. In their simulation, $c_0 = 0$, $c_e =1$ and $a_e$ was varied.

They considered the fixed domain case with 4 × 4 and 6 × 6 grids . They concluded that no estimator was uniformly superior for purposes of parameter estimation but the performance of all estimators was best when the spatial dependence was weak. In contrast, standard 95% prediction intervals performed best when the spatial dependence was strong with the MLE based interval having slightly better overall performance compared to the others. However, they concluded that very little would be sacrificed by using the more easily computed least squares estimators.

In semivariogram estimation, the asymptotic behavior in the expanding domain and fixed domain cases are different. Mardia and Marshall (1984) proved the consistency and asymptotic normality of the MLE of $(\sigma^2,\theta)$ in the expanding domain case for an underlying stationary Gaussian process Z and for more general isotropic semivariograms of the form

$$\gamma(h) = \sigma^2\rho(-h/\theta).$$

For the fixed domain case and a one-dimensional Gaussian process, Ying (1991) has shown that the MLEs for $\sigma^2$ and $\theta$ are not consistent individually but that the MLE for the ratio $\sigma^2/\theta$ is consistent and asymptotically normal. Stein (1999) indicated that this ratio is often more important for prediction than are the individual parameters. For the class of Matern models, of which the exponential model is a member, Zhang (2003) has shown that predicted values and prediction variances are approximately the same for models with different parameter values but the same ratio.

Webster and Oliver (1992) presented the results of a study which focused on the variability of the sample semivariograms at fixed lag distances for spherical and exponential models in the fixed domain case. They concluded that "semivariograms computed on fewer than 50 data are of little value and that at least 100 data are needed." Baczkowski and Mardia (1987) presented evidence via simulation that, for each fixed lag distance, the sample semivariogram from an underlying stationary Gaussian process in two dimensions with a spherical semivariogram is approximately lognormally distributed for moderately large sample sizes.

## 4. General Framework for the Simulations

As in the motivating example, we assumed that Z(x,y) was an isotropic, second-order stationary Gaussian process with E(Z) = 0. Two exponential models of the form (1) were considered. In both models, $c_0 = 1$ and $c_e = 4$. The models differed in the value of $a_e$, which was set to 2 or 3. Hence, for $a_e$ equal to 2 and 3, respectively, the values of the ratio $c_e/a_e$ were 2 and 4/3, the effective ranges were 5.5 and 8.3, and the correlations between grid points one unit apart were 0.48 and 0.57, respectively. The values of $a_e$ were chosen so that in the 10 × 10 grid the effective ranges were approximately ⅓ and ⅔ of the maximum distance between any pair of grid points (opposite corners of the grid).

Consideration was restricted to square grids with grid points equally spaced in both directions. In the expanding domain case, grid points were always spaced one unit apart. Four grid sizes were included: 7 × 7, 10 × 10, 14 × 14 and 20 × 20, corresponding to sample sizes of 49, 100, 196 and 400, respectively. For the fixed domain case, the overall grid size was fixed at 10 units in each direction and grid points were spaced at 1.5 units apart (7 × 7 grid), 1.0 units apart (10 × 10 grid) and 0.5 units apart (20 × 20 grid).

For each parameter combination, 1000 realizations for the 20 × 20 grid were generated using SAS's PROC SIM2D. The realizations for the smaller girds were obtained as subsets of the 20 × 20 grid anchored at the lower left corner of the grid; i.e., at (x,y) = (1,1). Sample semivariograms were calculated using PROC VARIOGRAM.

For each sample semivariogram, an exponential model was fitted using iteratively reweighted least squares where the weights were proportional to the reciprocal of the estimated variance of the semivariogram at that point. The algorithm was terminated if convergence was not obtained after 100 iterations. The nugget $c_0$ was restricted to be non-negative and $c_e$ and $a_e$ were restricted to be positive.

## 5. Simulation in the Expanding Domain Case

The results for both models were similar and only results for $a_e = 3$ are presented here.

As with any nonlinear model fitting, convergence of the algorithm can be problematic. The convergence status according to PROC NLIN for each grid size is summarized in Table 1. All realizations in which the Hessian was singular were excluded from further analysis. Zero estimates became less of a problem as the sample size increased. The zero estimates were for $c_0$ except for nine realizations in the 7 × 7 grid where $a_e$ was estimated to be zero. In three of these cases, $c_0$ was also estimated to be zero. However, in all cases the Hessian was singular and the realizations were automatically excluded from further considerations.

Figure 3 illustrates the variability in the sample semivariograms for the first 50 realizations for each grid size. From the figure it is clear that the sample to sample variability decreases as the sample/grid size increases. In addition, the sample semivariograms tend to become "smoother" as the sample size increases. Side-by-side boxplots of the sample semivariogram values at each observed lag distance (data not shown) show skewed distributions which appear to be consistent with the lognormality result of Baczkowski and Mardia (1987).

Figure 4 shows the weighted least squares fits for the first six sample semivariograms from the 10 × 10 grid where convergence was obtained. Realizations 2 and 4 from Figure 1 had singular Hessians and are not included. The figure represents the common situation for all grid sizes; viz., the fitted semivariograms were "close" to the sample semivariograms but often neither the sample nor fitted semivariogram were close to the underlying theoretical exponential semivariogram model.

Plots of the estimated coefficients which define the exponential semivariogram are presented in Figure 5. Boxplots for the individual coefficents and the ratio $c_e/a_e$ are shown in Figure 6. Both figures have truncated upper tails for $c_e$ and $a_e$ with 31 and 61 observations, respectively, over all four grids not plotted. The maximum values occurred for the same sample on the 14 × 14 grid where the estimates were 673.9 for $c_e$ and 1792.8 for $a_e$. The estimate for $c_0$ was 1.9 in that sample, which was slightly larger than the 90th percentile of its distribution. In general, realizations with large estimated values for $a_e$ tended to have large values for $c_e$.

The sampling distribution of $c_0$ is highly skewed in the 7 ×7 grid with a median of zero and an upper quartile of 1.002 (which was essentially the true value of 1.0). The skewness decreases as the grid size increases but is still present in the 20 × 20 grid. The downward bias in the estimates of $c_0$, caused, in part, by the large number of zero estimates, decreases as the grid size increases. The sampling distributions of $c_e$ and $a_e$ display similar skewed behavior but with much less pronounced bias. In contrast, the distribution of the ratio $c_e/a_e$ is relatively symmetric with an upward bias which decreases as the grid size increases.

## 6. Simulation in the Fixed Domain (In-fill) Case

The results for the fixed domain case are similar to the expanding domain results and are not presented here. Convergence problems were similar to those described previously and only realizations where convergence was obtained were included in the analysis. The only notable difference between the two cases is that the decreasing variability of the individual parameter estimates as the sample size increases (and the distance between grid points decreases) is not as pronounced in the fixed domain case. This is evident by comparing Figure 8 for the fixed domain case to Figure 6 for the expanding domain case.

## 7. Approximate Confidence Intervals

Approximate 95% large sample confidence intervals based on the normal distribution and the asymptotic variance estimates from PROC NLIN were constructed for all four parameters. The variance of the estimated ratio $c_e/a_e$ was approximated using the delta method. Coverage levels for the expanding domain case when $a_e = 3$ are given in Table 2 and confidence average interval lengths are presented in Table 3. Similar results hold for the other cases.

Empirical coverage levels for all parameters are well below the nominal level and decrease for both $c_e$ and $a_e$ as the sample/grid size increases. The increasing coverage for $c_0$ as a function of sample size is a reflection of the decreasing number of zero estimates for the parameter. When realizations with one or more zero estimates are removed from consideration, the coverage levels improve but are still well below the nominal level. In addition, in the reduced set of realizations, the coverages for all parameters decrease dramatically as the sample size increases.

For all parameters, the distribution of the confidence interval lengths is highly skewed with a long upper tail as evidenced by the large differences between the mean and median lengths in Table 3. The skewness decreases somewhat as the sample size increases. Average confidence interval length also decreases as sample size increases for all parameters.

A graph of confidence interval length versus the parameter estimate for the 20 × 20 grid is shown in Figure 7. The corresponding graphs for the remaining grid sizes show similar patterns. The figure provides some insight into the relationship between the coverage level and confidence interval length. For all parameters, except when the estimate is very close to the true value, the longer confidence intervals for a particular estimated value are those which covered the true value. Since larger standard errors arise from larger error sums of squares in the least squares fit, the realizations where the confidence intervals cover the true parameter values are those where there is more variability in the sample semivariogram; i.e., “smooth” sample semivariograms do not lead to “good” parameter estimates.

## 8. Conclusion

The results from our simulations lead to several conclusions about the ability to estimate the semivariogram in small samples assuming a correctly known functional form. First, there is huge sample to sample variability in the sample and fitted semivariograms. While this fact has been commented on the literature, its extent has not been completely clear. The most important practical ramification of this variability is the difficulty it causes a user in recognizing an appropriate model from a plot of the sample semivariogram for his/her data. For example, which, if any, of the realizations displayed in Figure 1 would have been fit to an exponential model if they represented real data?

The simulation results also demonstrate that weighted least squares generally provides a fitted semivariogram "close" to the sample semivariogram. Unfortunately, neither may be close to the true model. This is reflected in both the variability in the parameter estimates and in large confidence interval lengths and coverage levels well below the nominal level. While decreasing standard errors for estimated parameters and confidence interval lengths as sample sizes increase might be expected, the accompanying decrease in confidence interval coverage is troublesome, at least in the expanding domain case where estimators have been shown to be consistent. The issue is complicated by the relationship between coverage and confidence interval length.

Finally, the most disturbing impact of our results is with regards to our motivation for doing the study. For many agricultural field experiments, and perhaps for other disciplines as well, sample sizes of 100 and 200 may not be feasible, either practically or financially. Hence, from the experimenter's point of view, the "small" samples in our simulation are not really "small" samples at all.

## References

Baczkowski, A. J. and K. V. Mardia (1987). Approximate lognormality of the sample semivariogram under a Gaussian process. *Communications in Statistics: Simulation and Computation*, 16, 571-585.

Diggle, P. J., P. J. Ribeiro, Jr. and O. F. Christensen (2002). An introduction to model-based geostatistics. *In* Spatial statistics and computational methods. J. Møller, ed. New York: Springer Verlag, 43-86.

Mardia, K. V. and R. J. Marshall (1984). Maximum likelihood estimation of models for residual covariance spatial regression. *Biometrika*, 71, 135-146.

Stein, M. L. (1999). Interpolation of spatial data: Some theory for kriging. New York: Springer Verlag.

Webster, R. and M. A. Oliver (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, 43, 177-192.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36, 280-196.

Zhang, H. (2003). Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. *Journal of the American Statistical Association*. In press.

Zimmerman, D. L. and M. B. Zimmerman (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, 33, 77-91.

~~~~~~~~~~~~~~~

Table 1. Convergence status for realizations from the exponential model with $a_e = 3$ as determined by SAS's PROC NLIN in the expanding domain case.

| | Grid size | | | |
|---|---|---|---|---|
| Convergence status | 7 × 7 | 10 × 10 | 14 × 14 | 20 × 20 |
| Converged | 104 | 220 | 302 | 388 |
| No step size improvement | 225 | 289 | 381 | 450 |
| Zero estimate(s) | 393 | 331 | 248 | 154 |
| Singular Hessian | 263 | 159 | 69 | 8 |
| Singular Hessian & zero estimate(s) | 15 | 1 | 0 | 0 |
~~~~~~~~~~~~~~~

Table 2. Empirical coverage percentages for approximate large sample 95% confidence intervals for the parameters from the exponential model with $a_e = 3$ in the expanding domain case. Percentages in parentheses represent coverage levels without realizations having one or more parameters estimated to be zero.

| | Grid size | | | |
|---|---|---|---|---|
| Parameter | 7 × 7 | 10 × 10 | 14 × 14 | 20 × 20 |
| $c_0$ | 41.3 (90.5) | 46.0 (75.8) | 48.3 (65.8) | 52.2 (61.8) |
| $c_e$ | 53.6 (82.3) | 43.9 (59.6) | 37.1 (45.8) | 33.0 (37.2) |
| $a_e$ | 51.0 (96.0) | 52.7 (81.7) | 41.8 (54.2) | 30.5 (35.1) |
| $c_e/a_e$ | 52.4 (82.3) | 40.0 (61.6) | 41.8 (56.7) | 43.3 (51.4) |

Table 3. Average 95% confidence interval lengths for the parameters from the exponential model with $a_e = 3$ in the expanding domain case. Table entries are mean lengths with median lengths given in parentheses. Realizations with zero estimates are included.

| | Grid size | | | |
|---|---|---|---|---|
| Parameter | 7 × 7 | 10 × 10 | 14 × 14 | 20 × 20 |
| $c_0$ | 2.8 (0.0) | 2.9 (1.0) | 1.2 (0.9) | 1.0 (0.8) |
| $c_e$ | 230.8 (2.1) | 259.7 (1.7) | 196.8 (1.2) | 5.4 (0.9) |
| $a_e$ | 362.3 (2.6) | 616.0 (2.6) | 547.5 (1.8) | 16.1 (1.2) |
| $c_e/a_e$ | 9.5 (2.6) | 9.2 (1.5) | 1.6 (1.1) | 1.1 (0.8) |

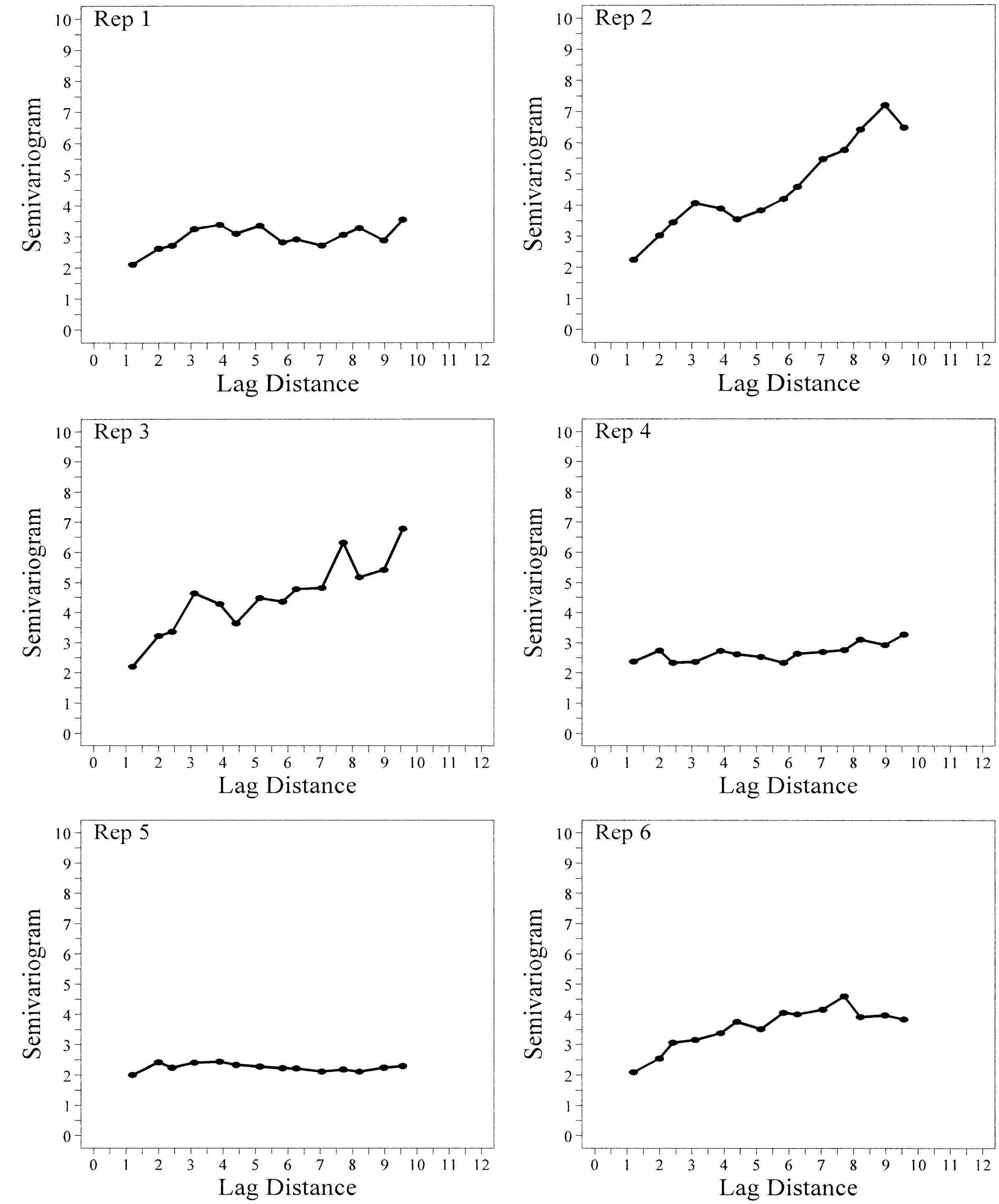

Figure 1. Sample semivariograms for the first six realizations from an exponential semivariogram model with $c_0 = 1$, $c_e = 4$ and $a_e = 3$.

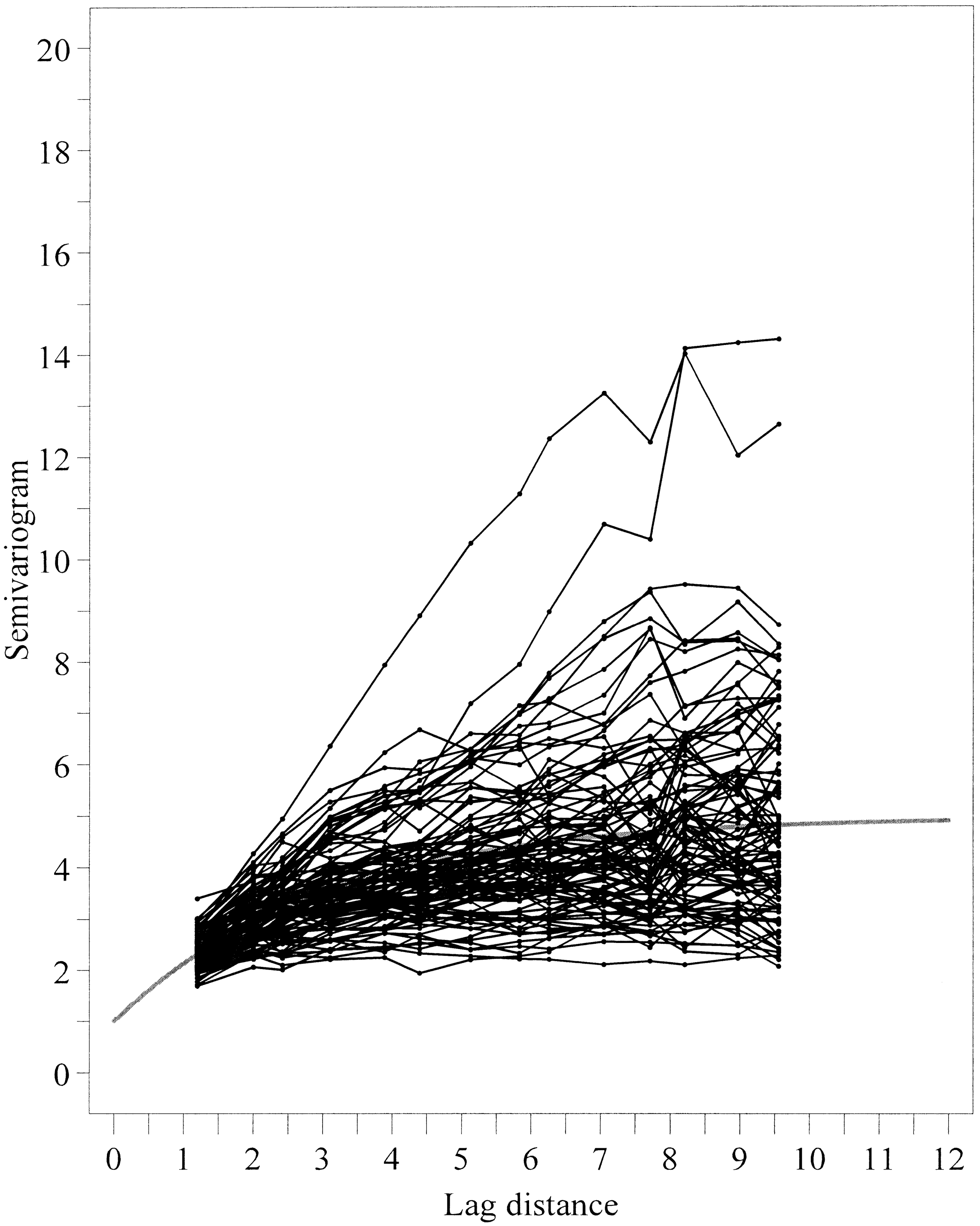

Figure 2. Exponential semivariogram model (solid line) with $c_0 = 1$, $c_e = 4$ and $a_e = 3$ on a 10 × 10 grid and 100 sample semivariograms.

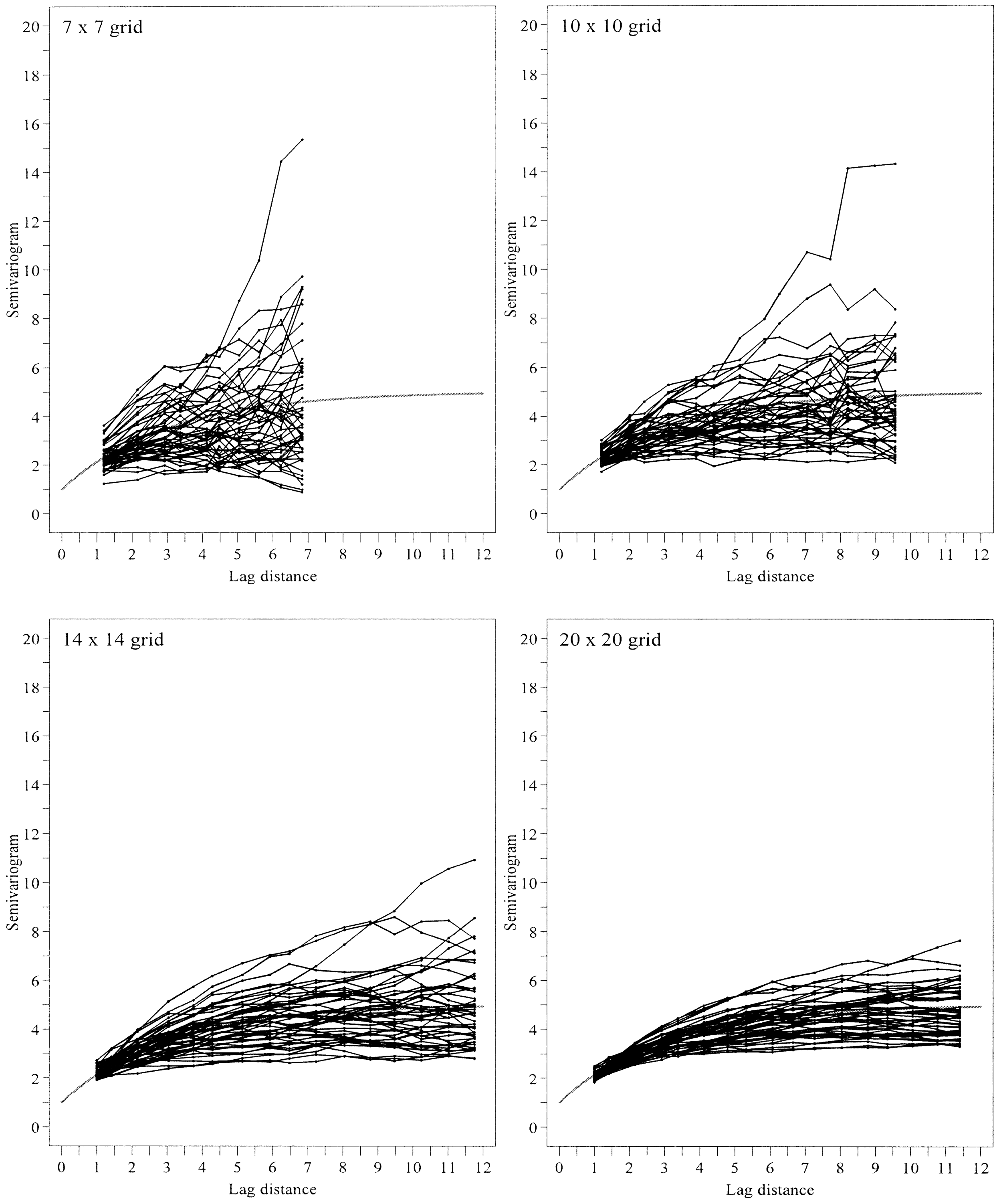

Figure 3. Exponential semivariogram model (solid line) with $c_0 = 1$, $c_e = 4$ and $a_e = 3$ and the first 50 sample semivariograms for each grid size in the expanding domain case.

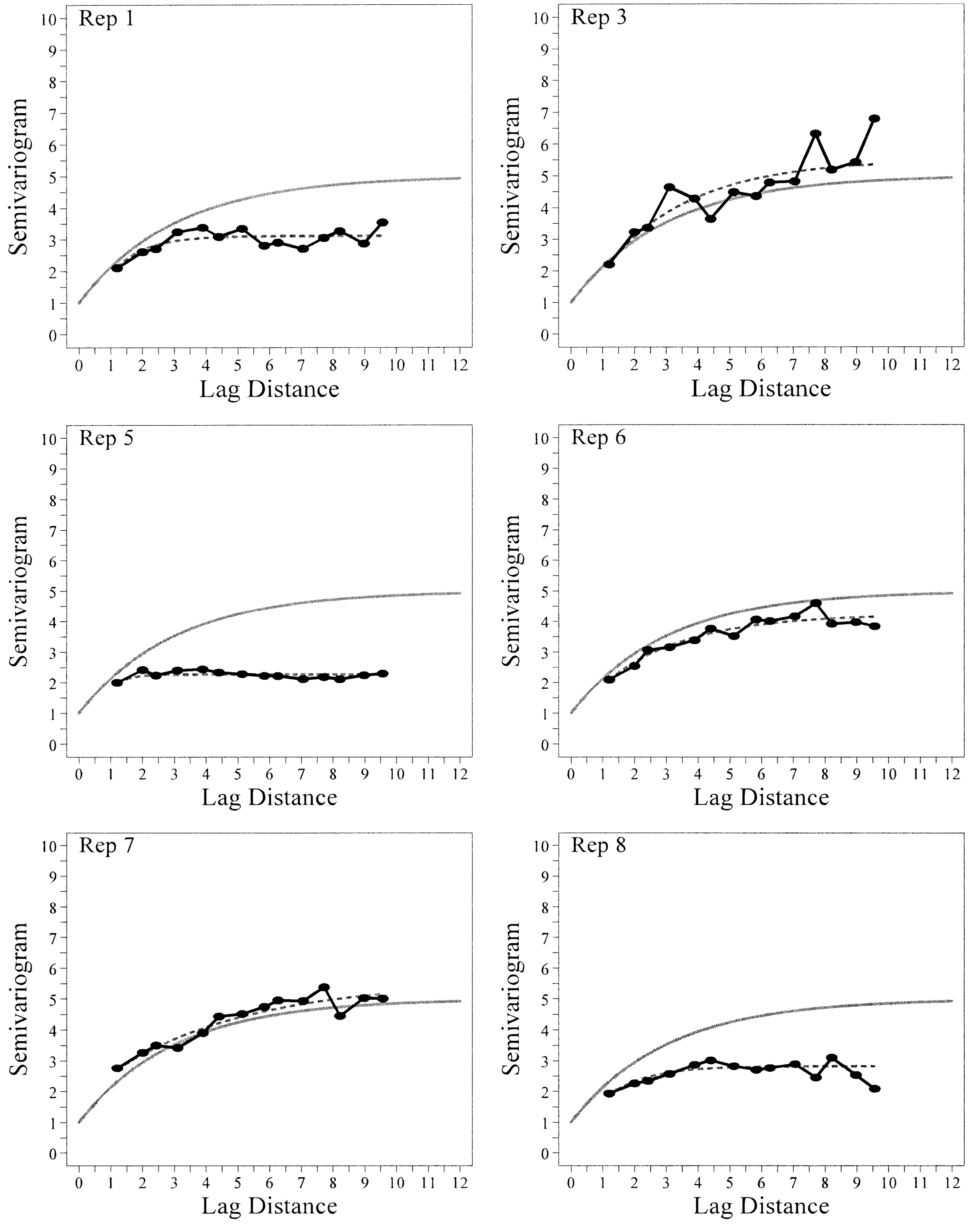

Figure 4. Sample semivariograms, weighted least squares fitted semivariograms (dotted line) and underlying exponential semivariogram model (solid line) with $c_0 = 1$, $c_e = 4$ and $a_e = 3$ in the expanding domain case.

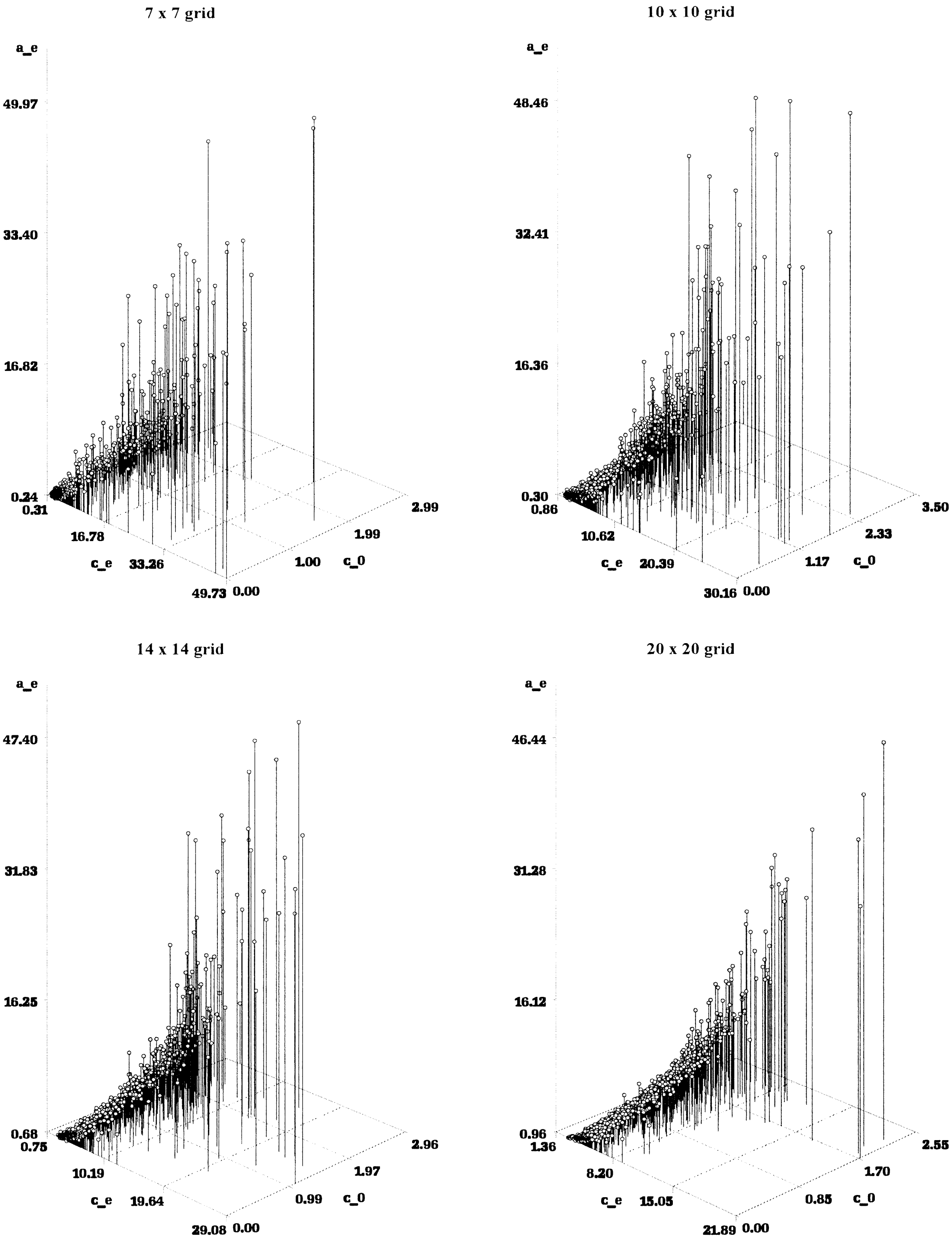

Figure 5. Estimated coefficients for an underlying exponential model with $c_0 = 1$, $c_e = 4$ and $a_e = 3$ in the expanding domain case. The scales on the axes for $c_e$ and $a_e$ have been truncated.

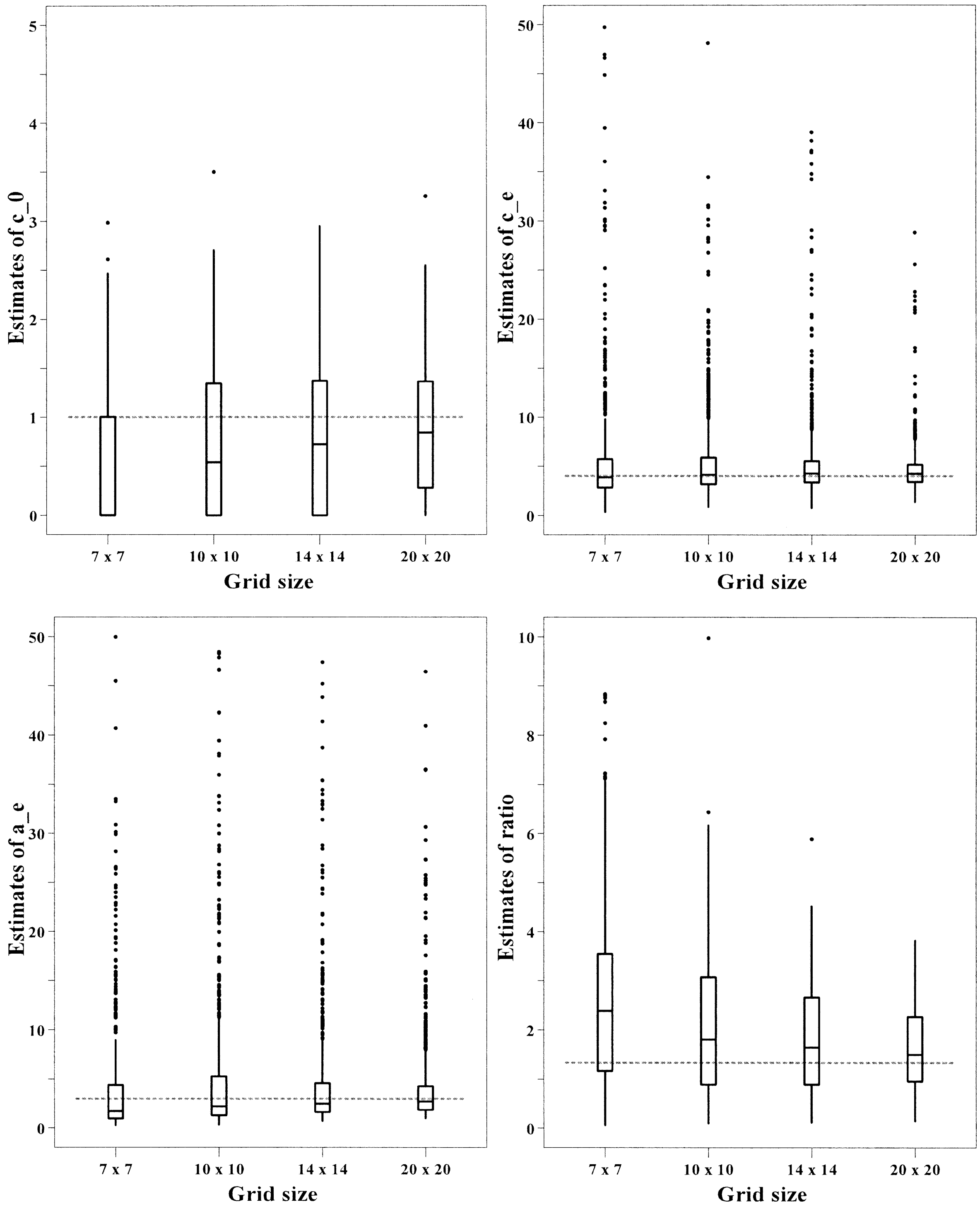

Figure 6. Sampling distributions for the estimated parameters of an underlying exponential model with $c_0 = 1$, $c_e = 4$ and $a_e = 3$ in the expanding domain case. The distributions for $c_e$ and $a_e$ have been truncated. Dotted lines represent the true parameter values.

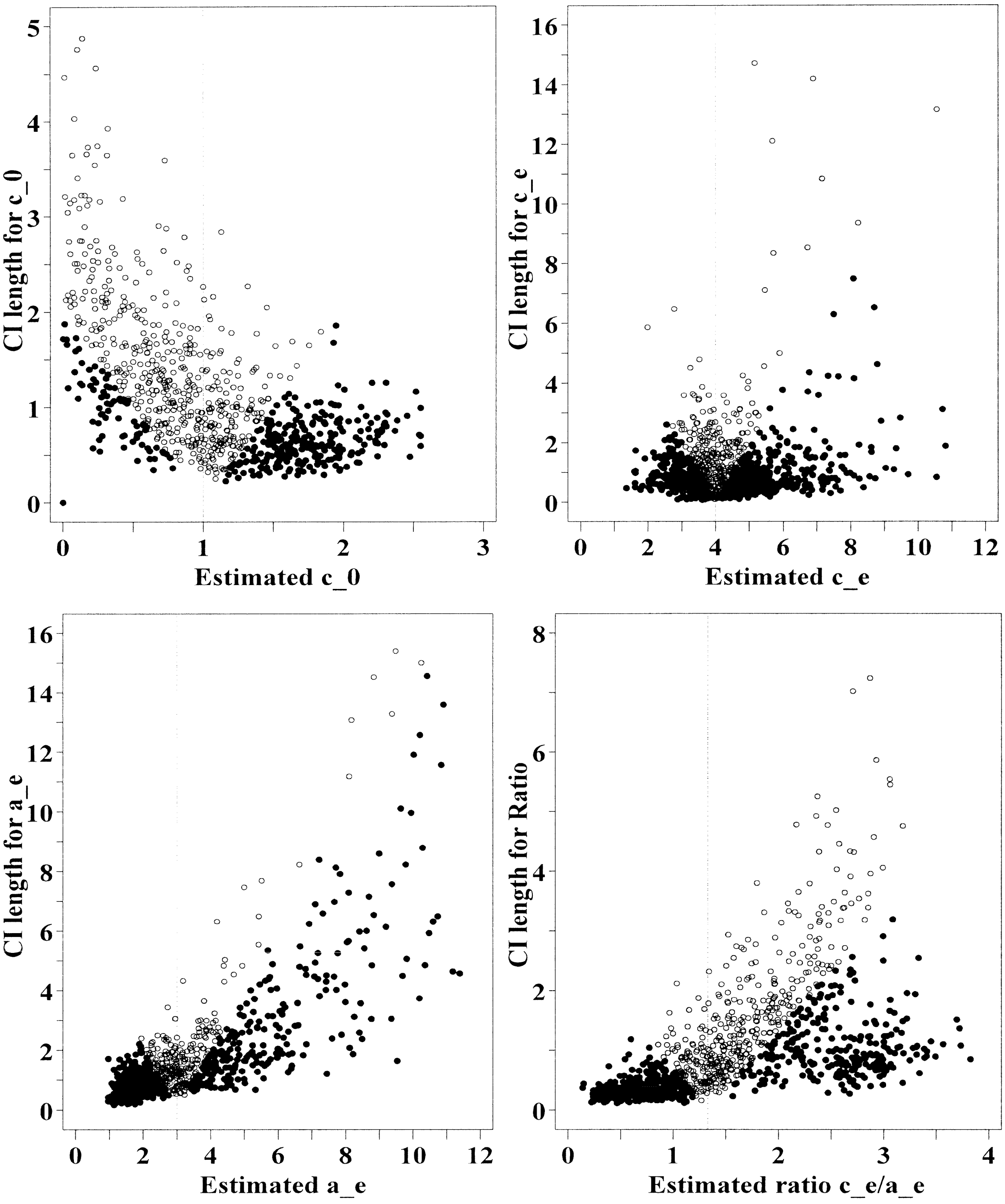

Figure 7. Estimated parameters versus confidence interval (CI) lengths for an underlying exponential model with $c_0 = 1$, $c_e = 4$ and $a_e = 3$ in the expanding domain case. A solid circle indicates the CI did not cover the true parameter value; an open circle indicates it did. "Outlying" data points for all four plots have been truncated. Dotted lines represent the true parameter values.

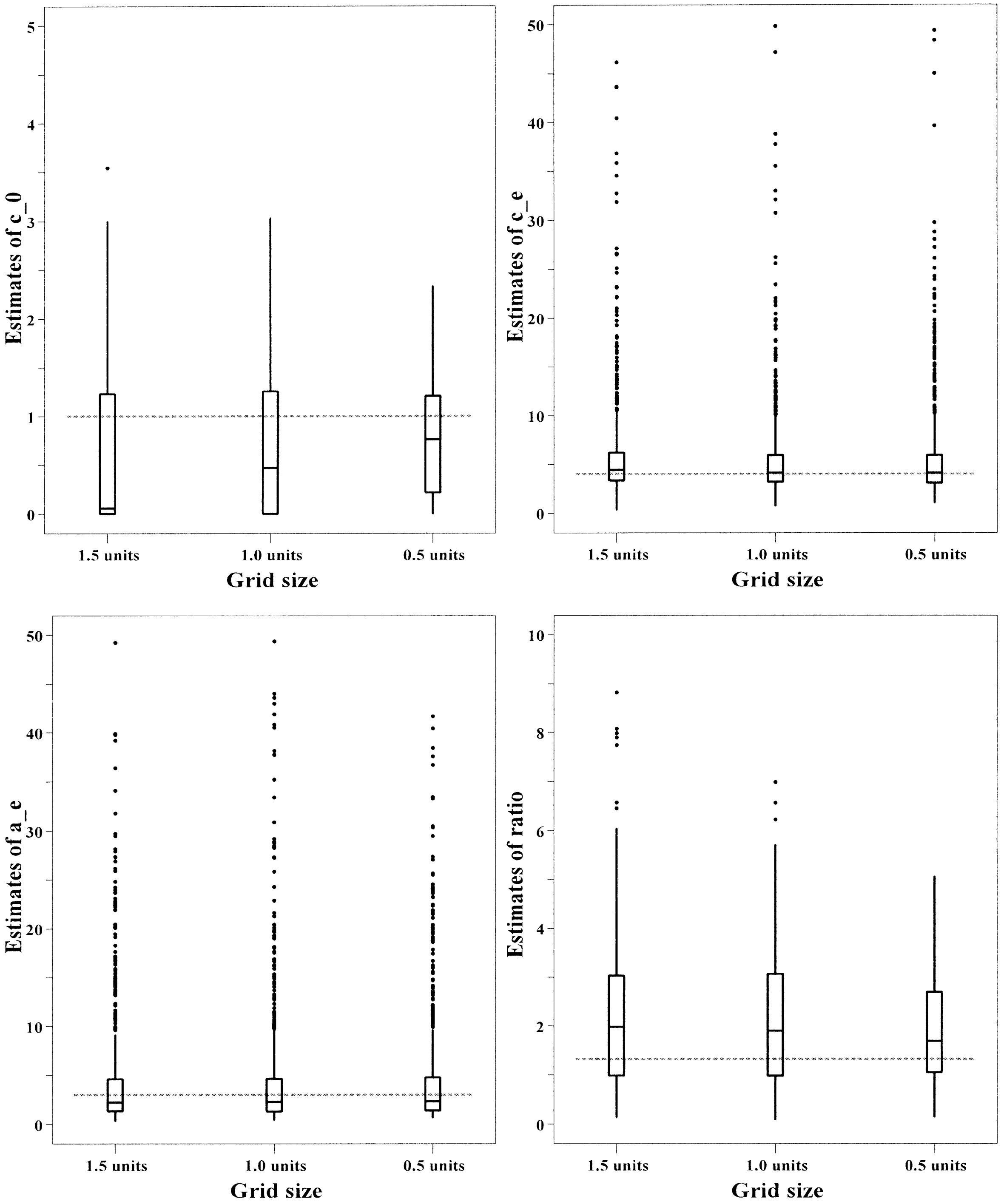

Figure 8. Sampling distributions for the estimated parameters of an underlying exponential model with $c_0 = 1$, $c_e = 4$ and $a_e = 3$ in the fixed domain case. The distributions for $c_e$ and $a_e$ have been truncated. Dotted lines represent the true parameter values. The 1.5 unit spacing is a 7 × 7 grid, the 1.0 unit spacing is a 10 × 10 grid and the 0.5 unit spacing is a 20 × 20 grid.